DOCUMENT RESUME

ED 403 896 IR 056 269

AUTHOR Eagan, Ann; Bender, Laura
TITLE Spiders and Worms and Crawlers, Oh My: Searching on the World Wide Web.
PUB DATE 96
NOTE 10p.; A downloaded version of one of 29 conference papers (for those in ERIC, see IR 056 258-274) placed on the Internet under the following overall title: "Untangling the Web. Proceedings of the Conference Sponsored by the Librarians Association of the University of California, Santa Barbara, and Friends of the UCSB Library" (Santa Barbara, CA, April 26, 1996).
AVAILABLE FROM Electronic version: http://www.library.ucsb.edu/untangle/
PUB TYPE Guides - Non-Classroom Use (055) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Access to Information; Information Needs; Integrated Library Systems; Internet; Librarians; *Online Searching; *Relevance (Information Retrieval); Research Needs; *Search Strategies; User Needs (Information); User Satisfaction (Information); *World Wide Web
IDENTIFIERS *Computer Searches; *Search Engines

ABSTRACT

Searching on the world wide web can be confusing. A myriad of search engines exist, often with little or no documentation, and many of these search engines work differently from the standard search engines people are accustomed to using. Intended for librarians, this paper defines search engines, directories, spiders, and robots, and covers basics for searching, providing criteria for choosing search engines as well as comparing some available search engines. Because the Internet is always growing and because search engines search in different ways and different parts of the Internet, doing the same search using different search engines will often produce widely differing results. Even yesterday's search will yield completely different results today. The concept of an expert as someone who knows almost everything about a subject is no longer valid. A better definition may be that an expert is someone who adapts to new information, digests it more quickly, and soon is hungry for more. A selected bibliography of articles on world wide web search engines is provided. (Author/SWC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Ann Eagan
Laura Bender
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Spiders and Worms and Crawlers, Oh My: Searching on the World Wide Web

ED 403 896

Ann Eagan
Science-Engineering/Undergraduate Services Librarian
Laura Bender
Science-Engineering/Undergraduate Services Librarian
University of Arizona, Tucson

Copyright 1996, Ann Eagan and Laura Bender. Used with permission.

Abstract

Searching on the world wide web can be confusing. A myriad of search engines exist, often with little or no documentation, and many of these search engines work differently from the standard commercial search engines we are used to using.

The workshop will begin with a guided search exercise. At the completion of the exercise, participants will be given a detailed information packet containing information on all the material to be covered during the session. We will then describe and demonstrate the use of several representative web search engines, explain some of the differences between web search engines, provide guided exercises for hands-on participation, and answer questions from the audience.

This workshop is aimed at librarians desiring to know how, when and why to search the Internet.

Searching on the world wide web can be confusing. A myriad of search engines exist, often with little or no documentation, and many of these search engines work differently from the standard commercial search engines we normally use. There are also many directories that attempt to organize the Internet by subject, and, today, there are many search engines that combine directory and keyword search capability. This paper will define search engines, directories, spiders and robots, cover some basics of searching, provide criteria for choosing search engines as well as a comparison of some of the search engines available.

Some caveats before we begin. There are dozens of search engines and several search engines for search engines, making it impossible to cover all of them. Also, much of what is written in this paper today is likely to be superseded by new information by the time you read it.

IR056349

BEST COPY AVAILABLE

What are Search Engines and Directories?

Search engines in use on the Internet use automated programs, called robots, to search the web. These automated programs are also known as spiders, crawlers, wanderers and worms. The robots crawl about the web indexing web sites. Some of them index web sites by title, some by uniform resource locators (URLs), some by words in each document in a web site, and some by combinations of these. Because the Internet is always growing and because these search engines search in different ways and search different parts of the Internet, doing the same search using different search engines will often give you wildly differing results.

Many directories on the Internet were created by humans tired of stumbling about the Internet looking for topics of interest. These personal lists grew in size and complexity, and eventually the humans started to use the available search engines to assist them in their quest to bring order to the mess. Yahoo is perhaps the best known of the directories. It was started by a couple of students at Stanford and now employs a variety of people, including librarians, who review and categorize web sites. Yahoo also now employs a search engine, as do most of the other directories. In addition, many of the search engines offer directories of topics for those who prefer to browse.

How to Search

Browsing a directory is a simple matter of following the links for the topic of interest. Searching either a directory or the portion of the web that a search engine covers works very much the same in almost all search engines. The basic format is that of a dialogue box, pane, or line where search terms can be entered followed by options to either submit or clear the search.

Once the search request is received, the search engine searches its own indexed database first, then, based on design, sends out spiders or other robots to add to the database. Results are sent back to the searcher, some annotated extensively, with links to the sources retrieved.

Full featured search engines also have options to expand or limit searches in a variety of ways. For example, in Lycos, the basic search assumes a boolean "or", which means that two or more terms will return results if any of the terms occur in documents indexed by Lycos. To obtain documents containing all the terms in a search, the Enhance Your Search option must be chosen and adjustments made to the default options.

Choosing a Search Engine

Choosing a search engine depends on the results you're looking for, though there are some criteria that may be useful. These criteria include:

Browsability -- how easy is it to understand the results? Do you receive enough information from the retrieved results to make a decision about the usefulness of the results?
Customizability -- can you construct a sufficiently detailed search so as to eliminate or greatly reduce irrelevant results?
Relevance -- no matter how browsable or customizable, are the results returned relevant to your search?

For example, searching for some information on the Native American squash blossom design using WebCrawler will bring relevant results, but either OpenText or InfoSeek would be better first choices because they both give more information to help you determine relevancy.

Comparing Search Engines

In the following chart, we have divided search engines into four categories: Classics, Leaders, Newer Kids on the Block, and Search Engines for Search Engines. By Classics we mean search engines that have been around for

awhile, that are well-known and well-used. Leaders are search engines that may or may not have been around for awhile, but are well-known, have high use and return relevant results. Newer Kids on the Block is our designation for more recent arrivals on the search engine scene. And Search Engines for Search Engines covers two meta-search tools that give you a single interface for searching multiple search engines at the same time. We will not be covering the collections of search engines such as Search.com, All-in-One, and CUSI(Configurable Unified Search Engine) that allow you to search different search engines in sequence.

The information given for each search engine is the name, the URL, how big the database is (if available), what it searches, general information on how to search, and why you might want to use it. Also included are characteristics specific to a given search engine. For example, MetaCrawler will check the links in the documents retrieved to ensure that they are valid, and OpenText allows you to see the keywords from your search in the context of the document.

Finding the information for the comparison chart was the result of an archaeological expedition -- a lot of digging in obscure places -- most of it on the help screens of the search engines themselves. OpenText is a good example of digging in obscure places: The help screen only shows up after you have done a search. The rest of the information comes from company information and the articles listed in the bibliography.

Comparison Chart--Search Engines

Classics

World Wide Web Worm: http://wwww.cs.colorado.edu/wwww/

A keyword oriented search engine good for general topic searching in a database of around 3 million sites. Has limited customization capability because it is forms based. Searches only http:// sites (no gopher, ftp sites). Once you make a connection to the server, the searching is very fast; of all the search engines we tried, though, this one took the longest to connect. The spiders in this engine seek out only URLs and web page titles for its index, so it's not the ideal place to find in-depth information on specific pages. You can search with "and" and "or" Boolean operators, and you can retrieve sound/graphics files. In fact, you can display GIF images from your search results--that's a plus. The down side is that there are no descriptions of the sites with the results. It searches strings, not words. That means all the terms in your query must appear in the order given to find a match. You can't bookmark sites, so searches have to be repeated.

So, why search with the Worm? It is good for simple, one or two-word topic searching, as well as generating lists of URLs in a certain area: Lists of business pages, organizations, etc.

WebCrawler: http://www.webcrawler.com

This database, now owned by America Online, has spiders that crawl over the entire web looking for popular sites. They index the contents of the documents as well as the URLs and titles, and claim to update their entire database of around 500,000 web pages on a monthly basis. There are no descriptions of the sites with the results, which makes gauging relevance difficult. However, in many simple, broad topic searches, relevant home pages appear at the top of the results list, allowing you to avoid scanning long lists of less relevant sites.

This engine searches for ftp and gopher sites, not just http's. It searches words, not strings. For example, a search for "colorado river" will turn up hits for those two words anywhere on the page. You can also search using the Boolean "and" and "or."

So, why search using WebCrawler? It's good for simple searches, has some customization capability--you can specify the number of words to search in your query and the number of desired results in blocks of 10, 25 or 100. You can also bookmark the results, making going back to specific sites very easy.

Yahoo!: http://www.yahoo.com

Open Text: http://www.opentext.com

One of the most popular sites to search: Hierarchical subject directory that merged with Open Text last November to add keyword searchability. Yahoo still has users contribute sites, but with the added capability of Open Text spiders, the database is scheduled to increase from 1.5 million pages to about 10 million pages--full text (this is supposed to happen any day now). Yahoo has a GUI (graphical user interface) that makes searching and browsing a piece of cake. It offers hourly news summaries from Reuters. Open Text search results are clearly marked, showing all URLs and the size of each. Results are scored by relevancy.

However, all these wonderful features of Open Text, including three types of searching, don't always work for simple queries. This is because the engine searches strings, not words. All words in a query must be present in the order given. However, the Boolean search capability is strong, and you can create your own weighted search. Yahoo! mixed with Open Text is a study in searching contrasts: On the one hand, the directory search does the work for you, on the other, you, the searcher, must do most of the work if you want the best results from the Open Text "power search."

So, why use Yahoo!? It's probably the best place to start any search of the Internet. It helps novices (and we're all novices in something) become acquainted with what the Internet has to offer.

EINet Galaxy: http://galaxy.einet.net/

The Galaxy is another hierarchical, topically organized search engine. Each topic has its own page in the Galaxy, and each page is organized into many lists. For example, the Topic List page provides links to other Galaxy pages containing specific information about your topic. Consists of a series of indexes from which to choose. For example, you can search an index of pages only found on the Galaxy itself, the web, gophers (to improve quality of gophers found, only those also referenced in Gopher Jewels appear in the index), Hytelnet--for access to thousands of telnet sites, and Galaxy Entries. This last index contains only information references in the Galaxy itself. Let's say you want to know if there are any references to the American Association of Retired People, or AARP. You can search on the full word or on the acronym to find out if you should continue your search further. Boolean "and," "or," and "not" can be used to refine the search process.

The Galaxy has a link "You can add information to this page!" Clicking on it will bring up a form which can be used to add references to an existing page, or send comments to Galaxy staff.

Each index provides its own results, which are scored according to the frequency specified keywords are found.

So, why use Galaxy? It allows the option of searching areas of the Internet not found on the web. It has a convenient browse page with preformatted searches on approximately 100 commonly chosen topics to save you time. Has topic lists and document lists relating to your topic.

Leaders

InfoSeek Guide: http://guide.infoseek.com

InfoSeek Guide is the free directory and keyword searchable service of InfoSeek. Use the Guide to direct your browsing of the Internet or to look for specific information. InfoSeek Guide indexes over 1 million web pages. It also indexes Usenet newsgroups, FTP and Gopher sites, e-mail addresses, and Frequently Asked Questions lists. Search features are many, and complex. But even with the complexity InfoSeek Guide offers great search

customizability and includes features such as: indexing of all words on a page, case sensitivity so that you can get a precise match on proper names, proximity searching, the "not" operator, symbol searching, and phrase searching. Results are ranked by relevancy and include that ranking, a link to the site of the information, the URL of the site, the size of the document, some description of the document, and a link to similar pages. You can bookmark your results too, making return visits to the sites much easier.

So, why use InfoSeek Guide? It's convenient (as of this writing it is the first search engine listed on Netscape's Net Search page) and offers many useful search features. Internet World tests also show it to provide the most relevant results (Venditto, 1996).

Lycos: http://www.lycos.com/

Back in December of 1995, Lycos claimed to have indexed 92% of the web. Now it claims to be the only complete guide to the Internet. Hype aside, they do have a huge database. They, too, have gone from being simply a keyword searchable index to adding a directory, which goes by the name of A2Z. Lycos also provides a service called Point, which provides reviews and ratings of the top 5% of all the Internet sites they index. Lycos searches every word in a web site and defaults, for some unfathomable reason, to an "or" search. To get the full range of search options you need to go into "Enhance your search". Once there, you can choose variations on "and" to match all your search terms, only two of your search terms or as many as seven search terms. You can also choose the level of relevancy of your search. The default is "loose match" which translates to a relevancy ranking of .1 on a scale where 1.0 is considered a perfect match. Display options range from showing 10-40 results per page in either standard, summary or detailed form. In the standard display includes a link to the document, the relevancy ranking, an outline, an abstract, the URL, and the size of the document.

So, why use Lycos? It covers a lot of the web, it is easy to use and the results are not only easy to read, but you also get enough information in the standard display to determine how relevant the results really are. You can also bookmark your results, making return visits much easier.

OpenText: http://www.opentext.com/omw/f-omw.html

OpenText provides little documentation on what or how it searches until you do a search, but it is popular because you do get results. The search form looks a bit intimidating at first, but is actually simple to use. You enter a word or phrase on each search line, indicate where you want to search (anywhere, summary, title, first heading, URL) and how you want to search (and, or, but not, near, followed by). Results include a link to the document, the relevancy ranking, the size, the URL, an excerpt describing the document, links to similar pages and an option to see the matches on the page. This option lets you see the key words in the context of the document.

So, why use OpenText? It offers a variety of sophisticated search options with a clear display of the results and extras such as links to similar pages and keywords in context. (See also Yahoo! under CLASSICS)

Newer Kids on the Block

Magellan: http://www.mckinley.com

Magellan offers added value to your searching by providing sites that have been evaluated by a staff of reviewers on the basis of depth, ease of use, and innovation. It also rates newsgroups, listservs and mailing lists.

You can search a directory mode: Explore Topics, or a keyword searchable mode: Search Magellan. Searches default to "or" if no other connectors are specified, and instructions are provided for Expanded Search utilizing more complex syntax.

Magellan provides a feature called Green Light which appears next to reviewed sites that, at the time of review, have no material "apparently intended for mature audiences." This feature pertains only to http sites, and applies only to the homepage itself, not to its links. The editors at McKinley make it very clear that a site with no green light is not necessarily "objectionable," it may simply contain topics the reviewers refer to as "adult."

Sites are ranked according to relevancy--frequency and proximity of your keywords in the results. The more relevant the site, the higher up on the list it's found.

So why use Magellan? Its spider uses natural language processing software to hunt down sites for the database. Although it's a small database, it's growing at a steady rate. Thousands of users submit their sites for review, and there are over 1.6 million unrated sites found by the Magellan robot awaiting review. Its value lies in the refereed sites and the ease of searching--both of which will improve with time.

Inktomi: http://inktomi.berkeley.edu/

Full-text search engine for the web that claims to be the fastest (1-2 second response time), and is named for a Trickster Spider of Plains Indian mythology that brought culture to the people. The Trickster also represents the weak vs. the strong, the triumph of the underdog. Inktomi will accept up to 20 words in a query, and ranks documents by how many of the search terms are found in it. The searcher is offered the option to display results with or without full graphics (dispensing with graphics could be a real time-saver). It also searches for same word roots instead of endings (e.g. watch, not watch-ing, or watch-ed). Using a + (plus) before a word indicates that it must be included in the results. A - (minus) indicates it must be excluded from the results.

Inktomi is a prototype project out of UC Berkeley--it will soon be moving into a commercial venture that will fully exploit its possibilities using leading edge equipment. It is based on parallel computing technology to build scalable web servers: Increase availability and automatically grow as the volume grows.

The scalability is incremental--the project is moving to what they call a 32-node version which brings with it the capability to handle approximately 100 million queries per week.

So, why use Inktomi? Because it represents the future of web searching. It may now provide too many irrelevant results, but the technology is improving and a new iteration is imminent.

Alta Vista: http://www.altavista.digital.com/

Alta Vista searches for words on web pages. It allows you to perform simple or complex searches and has speedy retrieval times and well-developed robot technology (spiders, etc.). If no connector is used in the search the default is "or." Truncation is possible, as are field searches in text, URLs, title and links. The link search retrieves pages where at least one link represented on that page matches your search query. Advanced searching is also available by using Boolean operators and adjacency symbols. The near symbol ~ can be used as can parentheses for nesting.

Web pages are evaluated for relevance--its ranking system is not as effective as that of other search engines because it indexes any and all references to a search term, no matter how far off it may be from the query's intent. Its search engine doesn't allow "stemming" as others do, which means that searches are performed only on the exact phrase--plurals and other forms of words are left out. However, if a document is found in your search, you can be sure your search terms are somewhere in it. Alta Vista also provides dates in its results list. Although you can refine your search by using the Power Search option, Alta Vista doesn't have as much on-screen help as other search engines. In terms of sheer scope, however, you'll know the Internet universe was scoured once your query is sent out. You can bookmark your results, making future site visits much easier.

So, why use Alta Vista? Because it searches for the obscure and hard-to-find subjects and performs its searches with speed. If you want to find as much as you can about a certain topic, this is the search engine for you. Its spider technology is powered by Digital's Alpha architecture, and claims to have 21 million, fully indexed pages in its database.

Excite: http://www.excite.com

This search engine offers two ways of searching: Concept or keyword. Many times there are no significant differences between the results of these searches. There is no Boolean searching, so trying to find specific information on a topic can be frustrating. The pluses of this engine, however lie in its service offerings: You can do a directory search, much like that of Yahoo!, or a keyword search. You can search for reviews, cartoons, news summaries, newsgroup texts and public ads. Unlike Alta Vista, its aim is not to build a comprehensive database, but one that is popular and current. The entire database is checked and updated weekly by spiders that are sent out on specific missions: One is sent to the What's New sites to compile a database of new URLs. Another is then sent out to bring back the page contents to the Excite database.

Excite took over as the site of choice for Netscape's Net Directory, replacing Yahoo! Perhaps the "one-stop-shopping" idea and the emphasis on currency and popularity had something to do with this decision.

There are some difficulties with the results displays. For example, you can't bookmark your results, so going back to check them can be a chore. There are no URLs displayed in the results either, making site visit choices harder. It is easy to use, however, and for current topics, a good place to start.

So, why use Excite? Because it incorporates the technology of the future: Concept searching, using natural language processing, needs to be further refined in this engine, but it's being utilized. Excite also provides a complete search service, with news, subject searching and classified ads.

Search Engines for Search Engines a.k.a. Meta Search Engines

MetaCrawler: http://metacrawler.cs.washington.edu:8080/

MetaCrawler is a search service that has no internal databases. It simply acts as a front end for 9 different search engines: OpenText, WebCrawler, Inktomi, Alta Vista, InfoSeek, Yahoo, Lycos, Excite, and EINet Galaxy. MetaCrawler sends your query to the search engines, then puts them into a uniform format for display. The search screen gives you a number of options. There is the usual search line but beneath it are 3 search options: search as a phrase (~3 min), search all these words (~ 1 min), search any of these words (~ 1 min). The times in parentheses indicate an estimate of the time it will take to complete the search. Below these search options are options to limit by regions of the world, by type of site, by the maximum amount of time you want to wait for results and by the minimum score. The results display returns the title of the document, selected text or an abstract (depending on the search engine), the relevancy ranking, the URL, and the search engine from which the information came.

So, why use MetaCrawler? It provides a single interface for 9 popular search engines, allows you to use some fairly sophisticated search options and will check the document URLs to make sure the link is valid.

SavvySearch: http://guaraldi.cs.colostate.edu:2000/form

SavvySearch is a search tool that provides a common interface for searching a variety of search engines. You enter your search on the Query line and it sends your query to multiple search engines. It ranks search engines by a number of factors, including how appropriate they might be and how fast the response time is currently. By requesting that the results be integrated, it will remove duplicate results! To search, enter the search words, choose the "and", "or", or "adjacency" operators from the query options, choose the number of results to be

returned from each search engine, choose the display format, tell it to integrate the results if you want, and wait. Since it is searching more than one search engine, the wait may be longer than that when using a single search engine. The normal display will give you most of the standard display for the specific search engine providing the results. If the results are coming from WebCrawler, you get the URL, if they are coming from OpenText, you will get the usual OpenText display. SavvySearch lists the name of the search engine providing the results. Another nice feature is that SavvySearch is currently available in 18 different languages.

So, why use SavvySearch? It's one stop shopping and it searches a lot of different search engines. In one search it reviewed 17 search engines as having possibly relevant information and searched 3 of them.

Conclusion

This is only a small portion of the ever-growing number of available search engines. There are many similarities and many differences in the way the search engines work. Think about what you want to get out of your search, try out a number of the search engines, and understand that the Internet and the search engines are changing daily. Yesterday's favorite search engine may be completely different today, and, most certainly, yesterday's search will provide completely different results today. The concept of an expert as someone who knows almost everything about a subject is no longer valid. A better definition may be that an expert is someone who adapts to new information, digests it more quickly, and soon is hungry for more.

A Definitely Not Comprehensive Bibliography of Articles on World Wide Web Search Engines

Compiled by Ann Eagan and Laura Bender

"Comparing Search Engines." [http://www.hamline.edu/library/links/comparisons.html].
A compilation of various articles on comparing search engines.

Decy, Don E. 1995. "All Aboard the Internet: Searching the World-Wide Web". *Techtrends* 40(4):7-8.
A good, overall explanation of searching and search engines on the Web.

Gralla, Preston. 1995. "Underground Internet." *PC Computing* 8(11):195-200+.
A compilation of information on browsers, multi-media viewers and players, and search engines.

Moeller, Michael. 1995. "Open Text, Yahoo Meld Search Engines." *PC Week* 12(38):135.
A brief article announcing the joining of OpenText and Yahoo.

Notess, Greg R. 1995. "Searching the World-Wide Web: Lycos, WebCrawler and More." *Online* 19:48-50+.
A good, general explanation of several popular search engines. The information is a little dated on Lycos. Also, the author favors traditional search techniques.

Scales, B. Jane and Elizabeth Caufield Felt. 1995. "Diversity of the World Wide Web: Using Robots to Search the Web." *Library Software Review* 14(3):132-136.
Another good overview of web search engines with information on how to choose a search engine. Some of the search engines covered are less well known than those covered in other articles. Also, while the article date is Fall 1995, much of the specific search engine information is out of date.

Udell, Jon. 1995. "Web Search." *Byte* 20(9):223-224+.
A techie article mostly on tools you can use to make your own information searchable. Contains a good quote:

"My advice to major Web contributors (and to creators of Web authoring tools) is to hire a library scientist".

Venditto, Gus. 1996. "Search Engine Showdown." *Internet World* 7(5):79-86.
A recent article comparing seven Internet search engines in lay terms. The results of Internet World's tests show InfoSeek Guide providing the most relevant results and Alta Vista the most comprehensive results.

Direct questions or comments about this page to webmaster@library.ucsb.edu .

Search "Untangling" Papers | "Untangling the Web" Home | InfoSurf Home

This is an official University of California, Santa Barbara Library web page.